Information Retrieval, 7, 73–97, 2004

Character N-Gram Tokenization for European Language Text Retrieval

PAUL McNAMEE mcnamee@jhuapl.edu
JAMES MAYFIELD mayfield@jhuapl.edu
Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723-6099, USA

Received December 5, 2002; Revised April 17, 2003; Accepted May 6, 2003

Abstract. The Cross-Language Evaluation Forum has encouraged research in text retrieval methods for numerous European languages and has developed durable test suites that allow language-specific techniques to be investigated and compared. The labor associated with crafting a retrieval system that takes advantage of sophisticated linguistic methods is daunting. We examine whether language-neutral methods can achieve accuracy comparable to language-specific methods with less concomitant software complexity. Using the CLEF 2002 test set we demonstrate empirically how overlapping character n-gram tokenization can provide retrieval accuracy that rivals the best current language-specific approaches for European languages. We show that $n = 4$ is a good choice for those languages, and document the increased storage and time requirements of the technique. We report on the benefits of and challenges posed by n-grams, and explain peculiarities attendant to bilingual retrieval. Our findings demonstrate clearly that accuracy using n-gram indexing rivals or exceeds accuracy using unnormalized words, for both monolingual and bilingual retrieval.

Keywords: cross-language information retrieval, language-neutral retrieval, character n-grams, Cross Language Evaluation Forum, European languages

1. Introduction

When designing a system for multilingual text retrieval, there appears to be a tradeoff between software complexity and the ability to operate over disparate languages. To obtain high-quality retrieval accuracy, many choose to incorporate language-specific resources, both for processing text in a single language and for translating documents or queries to another language for the purposes of retrieval. For example, IR systems typically utilize stopword lists, phrase lists, stemmers, decompounders, lexicons, thesauri, part-of-speech taggers, or other linguistic tools and resources to facilitate retrieval. Obtaining and integrating such resources is time-consuming and may involve considerable financial expense if commercial toolkits are relied upon. It is also possible that errors introduced by such tools offset much of the benefit gained. In this paper we present quantitative results that demonstrate how high-caliber retrieval accuracy can be obtained without language-specific resources. The basis of our approach is the use of overlapping character n-grams as indexing terms. This approach will easily scale to 20 or more languages—a requirement likely to be necessitated by EU enlargement.

We report on experiments using the Cross-Language Evaluation Forum (CLEF) test collections to evaluate retrieval accuracy in eight European languages. CLEF is unique among formal IR evaluations in creating a test collection with documents in so many languages; this makes it an ideal environment for language-neutral investigations into retrieval methodologies.

We first describe the historical use of character n-gram indexing and our approach to implementation and evaluation of n-gram-based retrieval. In the following section we report on monolingual experiments using n-grams for retrieval in all eight languages used in the CLEF 2002 evaluation. We then discuss issues that arise with bilingual retrieval over those same languages. We conclude with a description of the impact of n-gram use on performance, and some final remarks.

2. History

The use of character n-grams in language modeling dates back at least to Claude Shannon. While he is widely known for his juggling machinery (1980), Shannon also made contributions to the area of information theory. In his seminal paper (1948), Shannon described a sequence of character n-gram and word n-gram approximations to English.

The application of n-grams to information retrieval derived from the desire to decrease dictionary size. While the number of words that may be found in a collection is in theory infinite as the collection grows, the number of n-grams is bounded by $|alphabet|^n$. For small n, this number is quite tractable; when $n = 3$ for example, for the English alphabet of 26 letters plus space, at most 19,683 3-grams may be found. Thus, if memory constraints are severe, short character n-grams offer an attractive representation for the retrieval system's dictionary (which, unlike postings, is typically kept in memory).

With this goal in mind, numerous studies examined the efficiency of short, word-internal character n-grams. As early as 1974, de Heer (1974) explored the use of "n-polygrams" as an alternative to words. He termed the collection of n-grams that are derived from a word the *syntactic trace* of that word. Subsequent work (Willett 1979, De Heer 1982, Mah and D'Amore 1983, D'Amore and Mah 1985, Teufel 1988, Comlekoglu 1990) gradually increased n-gram length, studied varying length n-grams to homogenize term frequency, and increased test collection size.

In the 1990s, a shift occurred in how n-grams were viewed within information retrieval. Technical changes included an increase in n, and a shift to word-spanning n-grams. Qualitatively, these changes reflected a new view of n-grams as indexing terms in their own right, as opposed to simply serving as an indirect representation of words. These changes were hinted at in Cavnar (1994), and firmly established by Damashek (1995).

Reaction to the Damashek work (Harman et al. 1995) pointed out that Damashek's system did not perform up to the level of most other systems that participated in TREC-3. However, in addition to using n-grams as indexing terms, Damashek's system also used a novel similarity metric. The effects of these two technologies were not separated, so one cannot safely conclude from these results that n-grams are fundamentally inferior to words as indexing terms, even for the TREC-3 test set. In fact, throughout the early history of

n-grams as indexing terms, little distinction was made between the impact of n-grams and the impact of the particular similarity metric in use. This was understandable when the stated purpose for using n-grams was memory efficiency, but it makes little sense when trying to understand how n-grams affect retrieval accuracy.

Overlapping sequences of characters have been used for many applications other than document retrieval, including language identification (Cavnar and Trenkle 1994), spelling error detection (Zamora et al. 1981), keyword highlighting (Cohen 1995), and restoration of diacritical marks (Mihalcea and Nastase 2002). N-grams have been recognized for their ability to retrieve documents that have been degraded due to OCR errors (Pearce and Nicholas 1996). However, the largest application of character n-grams in information retrieval is probably in retrieval of Asian languages (Chen et al. 1997, Lee and Ahn 1996, Ogawa and Matsuda 1997). Written languages such as Chinese and Japanese do not include word separator characters. Therefore, a word-based approach to indexing demands a segmenter that can identify word boundaries. Not only is such a segmenter language-specific (requiring new training for each language to be segmented), its errors can also degrade the quality of the index. N-grams, in contrast, do not treat word separators as special in any way, and so proceed blissfully onward irrespective of whether separators are present.

3. Approach

The experiments presented herein were carried out using the Hopkins Automated Information Retriever for Combing Unstructured Text (HAIRCUT) system (Mayfield et al. 2000). HAIRCUT is a Java-based text retrieval engine developed at the Johns Hopkins University Applied Physics Laboratory. We are particularly interested in language-neutral techniques for HAIRCUT because we lack the resources to do significant language-specific work.

HAIRCUT has a flexible tokenizer that supports multiple term types. All text is read in as Unicode, using Java's built-in Unicode facilities. For European languages, the tokenizer is typically configured to break words at spaces, downcase them and remove diacritics. Punctuation is used to identify sentence boundaries, then removed. Stop structure is then optionally removed. We manually developed a list of 459 English stop phrases to be removed from queries. Each phrase was then translated into the other supported languages using various commercial machine translation systems. We lack the resources to verify the quality of such non-English stop structure, but its removal from queries seems to improve accuracy.

We use the resulting words, which we call *raw words,* as the main point of comparison with n-grams in this study. They also form the basis for the construction of n-grams. A space is placed at the beginning and end of each sentence, and between each pair of words. Each subsequence of length n is then generated as an n-gram. A text with fewer than $n - 2$ characters generates no n-grams in this approach. This is not much of a problem for 4-grams, but 6-grams are unable to respond, for example, to the query 'IBM.' A solution is to generate an additional indexing term for each word of length less than $n - 2$; however, this is not part of our ordinary processing.

Outside of the character-level processing required by the tokenizer, and the removal of our guesses at stop structure, HAIRCUT has no language-specific code. On occasion we have run experiments using one of the Snowball stemmers (Porter 2001), but this is not a regular part of our processing. Nor do we do any decompounding, lemmatization, part-of-speech tagging, chunking or parsing.

HAIRCUT uses a unigram language model similarity metric (Ponte and Croft 1998, Miller et al. 1999, Hiemstra 2000) with Jelinek-Mercer smoothing (1980). In this model, relevance is defined as

$$P(D \mid Q) = \prod_{q \in Q} [\alpha P(q \mid D) + (1 - \alpha) P(q \mid C)],$$

where Q is a query, D is a document, C is the collection as a whole, and α is a smoothing parameter. The probabilities on the right side of the equation are replaced by their maximum likelihood estimates when scoring a document. The language model has the advantage that term weights are mediated by the corpus. In addition to relieving the developer of the burden of identifying a term weighting scheme, this feature admits the potential for improved performance with a larger corpus.

This language model assumes that all query terms are independent. This is untrue for words, but wildly untrue for n-grams (adjacent n-grams share all but one letter). Nonetheless, the metric does not appear to suffer for its unrealistic assumption, even when applied to n-grams. The one effect that this increased level of dependence appears to have is to decrease the optimal value of the smoothing parameter α.

Our early tests on the language model had it consistently outperforming both Okapi BM25 (Robertson et al. 1999) and cosine (Salton and Buckley 1988) (although the differences may not have been statistically significant). The other methods seem to work reasonably well with n-grams too (e.g., Savoy 2003). It is possible that different similarity metrics respond better to different n-gram lengths, but this is speculation.

The HAIRCUT index is a typical inverted index. The dictionary is stored in a compressed B-tree, which is paged to disk as necessary. Postings are stored on disk using gamma compression (Witten et al. 1999). We keep only term counts in our postings lists; we do not keep term position information. We also store a bag-of-words representation of each document on disk to facilitate blind relevance feedback and term relationship discovery.

Blind relevance feedback for monolingual retrieval, and pre- and post-translation expansion for bilingual retrieval, are accomplished in the same way. Retrieval is performed on the initial query, and the top retrieved documents (typically 20 of them) are selected. The terms in those documents are weighted either according to the chi-squared statistic or according to our own proprietary affinity statistic $(L_i - C_i) \times IDF_i^K$, where L_i is the local frequency evaluated over the retrieved document set, C_i is the collection frequency, IDF_i is the inverse of the log of the document frequency, and K is a suitable constant. The highest weighted terms (typically 50 of them) are then selected as feedback terms.

The tests in this paper were performed on the CLEF 2002 collection. There are up to fifty queries per language in this collection, although some languages have fewer (not every topic has relevant documents in every language). We use mean average precision to measure retrieval accuracy. See the article in this volume by Braschler and Peters for details about the CLEF test collections (Braschler and Peters, 2004). HAIRCUT has consistently placed among the top systems in the CLEF monolingual and bilingual evaluations, giving some weight to the conclusion that n-grams rival today's language-specific approaches to tokenization.

4. N-grams for monolingual retrieval

Most systems that attempt retrieval of English texts use only light morphological normalization. The most frequently performed operation is suffix removal. A variety of stemmers attempt to reduce words to a canonical form for indexing; some are based on rules crafted for each language (Porter 1980, 2001) others rely on statistics of large corpora (Oard et al. 2001). Automated stemming faces the twin risks of conflating semantically unrelated words such as 'physicist' and 'physician,' and of leaving some related forms distinct.

Like stems, individual n-grams can be highly ambiguous. For example, the 4-gram 'mini' could be generated from 'dominion,' 'feminist,' 'miniature,' or 'ministry,' among others. While in isolation a single n-gram is ambiguous, n-grams can make up in volume what they lack in specificity. Thus the concept underlying a set of n-grams is often clear. For example, if the 4-grams 'prim,' 'rime,' 'mini,' 'inis,' and 'nist' occur in the same passage, very probably the phrase 'prime minister' was present. There is of course a chance that these n-grams were produced from a sentence like "The foreign minister ate prime rib for lunch;" however, the same ambiguity is faced by a word-based system. The redundancy afforded by n-grams is paid for with an increase in the size of the inverted index; a passage of k characters contains $k - n + 1$ n-grams of length n, but only approximately $(k + 1)/(l + 1)$ words, where l is the average word length for the language. As n increases in length, individual terms become both rarer and more determinate.

To see that n-grams can sometimes be less ambiguous than words, consider the common English phrase, 'white house'. In Table 1, we show frequency statistics for the 7-grams produced from 'white house' using the LA Times subset of the CLEF 2002 collection.

The word 'white' occurs 12308 times in the collection (as does the 7-gram '-white-') and the word 'house' occurs 15049 times. There are 4799 documents that contain both terms;

Table 1. Frequency of 7-grams produced for 'white house.'

Term	Document frequency	Collection frequency	IDF
-white-	12308	24916	3.163
white-h	3202	7286	5.106
hite-ho	2805	6819	5.297
ite-hou	2676	6662	5.365
te-hous	2845	6873	5.276
e-house	8706	18425	3.663
-house-	15049	34413	2.873

Document frequency is the number of documents in the collection that contain the term; collection frequency is the number of occurrences of the term in the collection; and inverse document frequency (IDF) is the log of the number of documents in the collection divided by the document frequency. Here, and throughout the paper, dashes are used in the printed representation of individual n-grams to indicate space characters.

however, the phrase ‘white house’ only occurs about 2660 times. A 7-gram like ‘ite-hou’ is much less common than ‘-white-’ and is a reliable indicator of the phrase ‘white house.’

4.1. Selecting N-gram size

The most common practice when using n-grams to represent text is to fix a single length of n. The length of n that maximizes retrieval quality varies by language. Long lengths of n increase lexicon size and will not represent short words well. For example the representation of abbreviations such as EU and US will be spread across a large number of 7-grams.

To ascertain suitable n-gram lengths for different languages, we ran the CLEF 2002 topics against the individual language collections, using multiple indexing term types. We considered n of lengths 3–7 as well as word-based indexing. Both title and description fields were used for a single pass of retrieval, but we did not use blind relevance feedback (Harman 1992). Blind relevance feedback would improve these results, and we did use this technique to good effect in our official submissions to the CLEF workshops. However, here we wanted to investigate the effects of different n-gram lengths in isolation. HAIRCUT uses a statistical language model for its similarity metric so we varied the smoothing parameter, α, from 0.10 to 0.90 for each indexing method. For a treatment of smoothing techniques for the language modeling approach to retrieval, see Zhai and Lafferty (2001).

Accuracy was measured using mean average precision as the metric. The maximal value for each length of n was obtained (often with different α values for different lengths of n). Charts of the accuracy of each indexing method are shown in figure 1.

We draw several conclusions from these graphs. First, it appears that $n = 4$ or $n = 5$ is best under the language model for many European languages. This comes as some surprise to us, since we have been using 6-grams for several years with apparent success; many of our official CLEF submissions in 2000, 2001, and 2002 were based on 6-gram indexing (e.g., McNamee et al. 2001a).

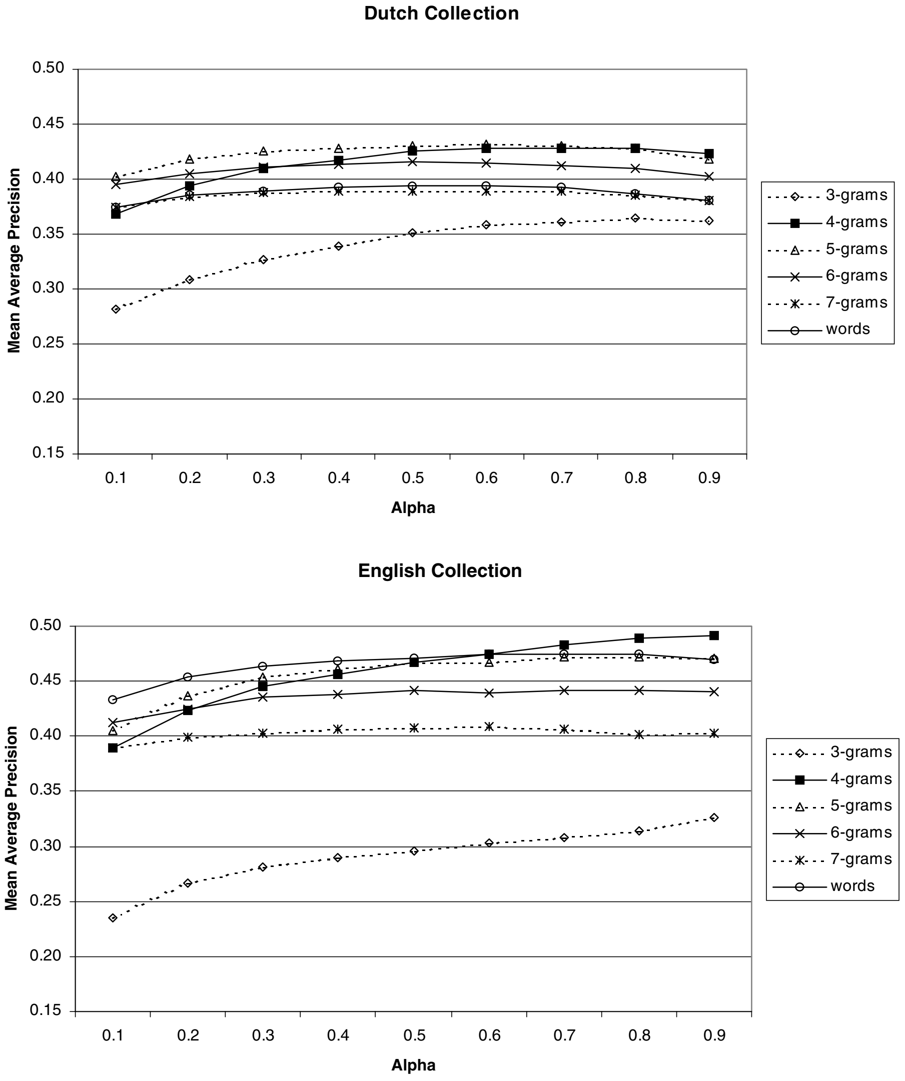
The maximum performance for each language is indicated in Table 2.

Table 2. Average precision using 4-, 5-, and 6-grams for each language using the optimal value for smoothing.

	4-grams		5-grams		6-grams	
	MAP	α	MAP	α	MAP	α
Dutch	0.4284	0.7	**0.4323**	**0.6**	0.4163	0.5
English	**0.4917**	**0.9**	0.4718	0.8	0.4420	0.7
Finnish	0.3839	0.5	**0.3968**	**0.3**	0.3760	0.2
French	**0.4097**	**0.7**	0.3899	0.7	0.3605	0.5
German	**0.4055**	**0.9**	0.4013	0.6	0.3782	0.7
Italian	**0.4051**	**0.7**	0.4041	0.6	0.3805	0.6
Spanish	**0.4678**	**0.7**	0.4479	0.5	0.4069	0.7
Swedish	**0.4312**	**0.8**	0.4116	0.6	0.3827	0.2

Maximal values are highlighted. The best performance using 3- or 7-grams was worse. All differences between 5-grams and 6-grams are statistically significant at the 0.05 level; differences between 4-grams and 5-grams are only significant for Swedish.

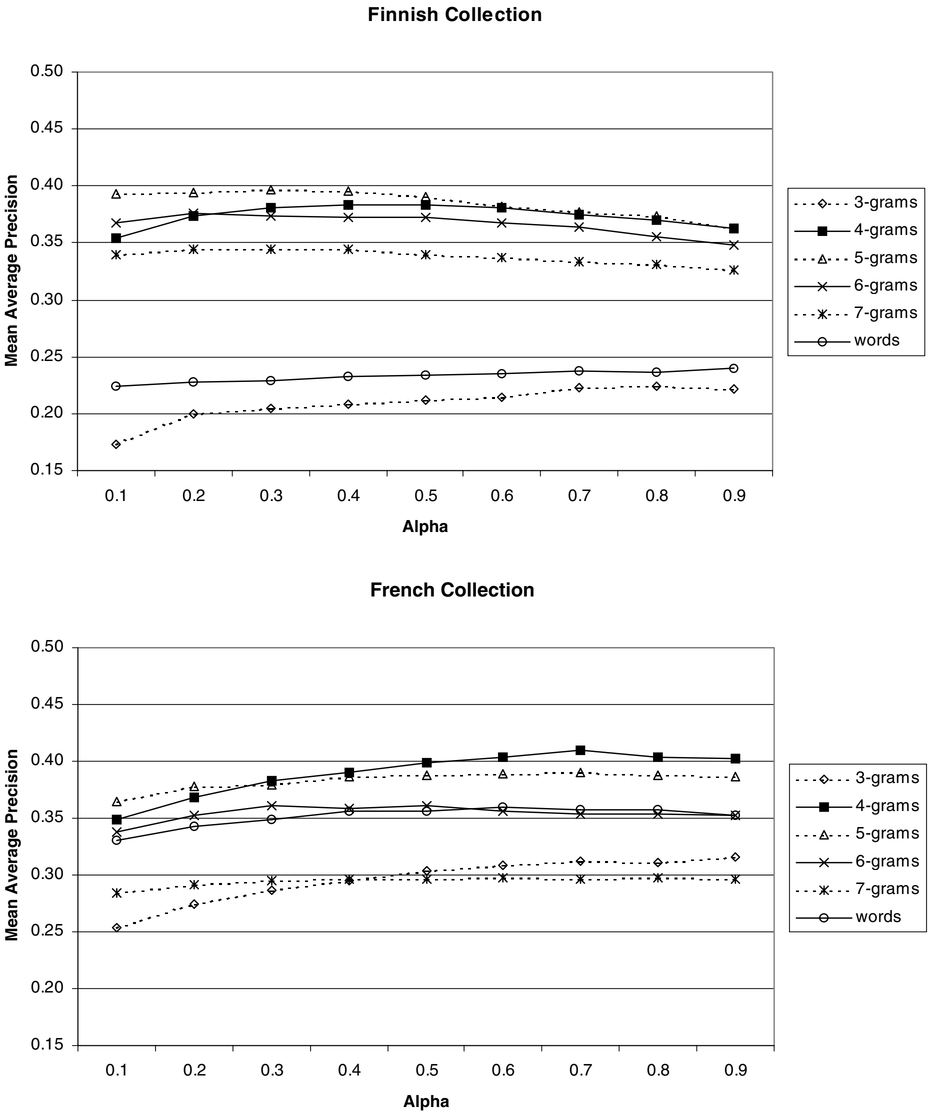
Higher values of α appear to be better when less discriminating terms, like 4-grams, are used. In effect, increasing α increases the scores of those documents that have a majority of the query terms. Figure 1 also indicates that the differences in accuracy due to the smoothing parameter are small. In particular, for a given indexing scheme, the differences in average precision as a function of smoothing constant are small, much smaller than differences due to the choice of indexing term type.



(a)

Figure 1. Comparative efficacy of n-gram indexing for eight languages.

(*Continued on next page.*)



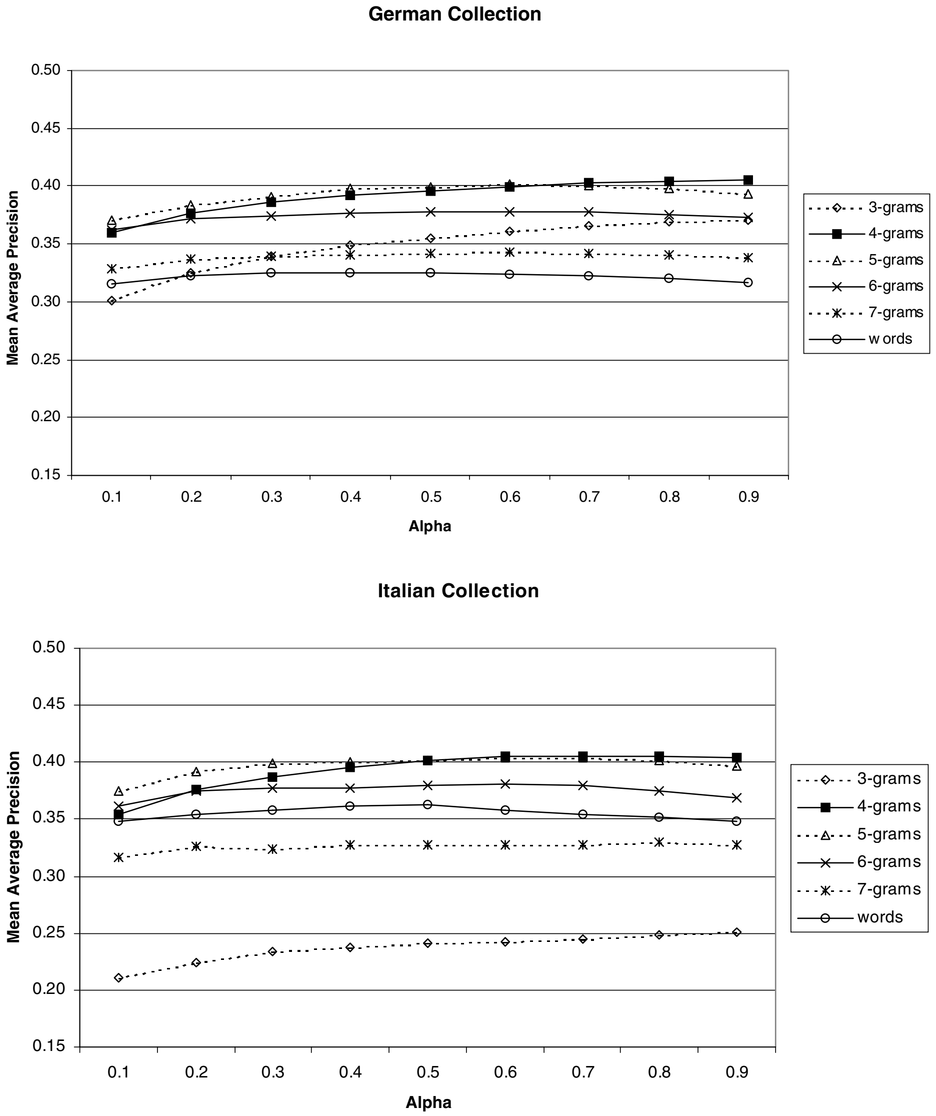
Finnish Collection

French Collection

(b)

Figure 1. (*Continued*).

We emphasize that these runs were created using only a single retrieval pass in a monolingual setting; different scenarios may require different lengths of n for optimality. Optimizing n under relevance feedback was beyond the scope of the present study, but is work that should be performed. We have started exploring methods for applying relevance feedback with n-grams as terms and our preliminary results suggest that longer lengths of n make better expansion terms, since longer n-grams are less ambiguous.

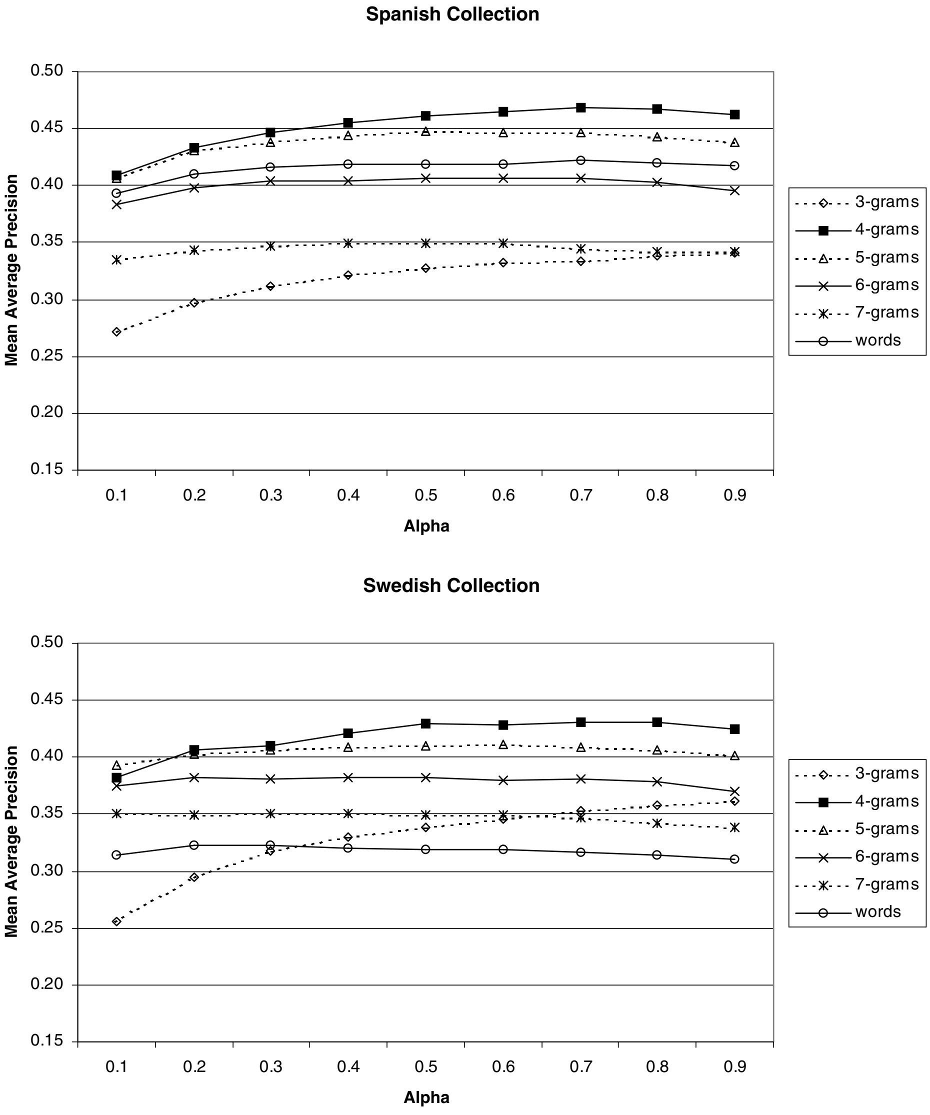


(c)

Figure 1. (*Continued*).

4.2. *N-grams versus words*

Another conclusion we draw from figure 1 is that *n*-gram indexing rivals or outperforms raw word-based indexing in each instance. For the more morphologically complex languages, *n*-grams show a decisive advantage over raw words. It is likely that decompounding,



(d)

Figure 1. (*Continued*).

stemming, or other normalization will result in better accuracy than the naïve use of raw words. We did not explore this because we lack the resources to do significant language-specific processing across the range of CLEF languages.

Table 3 lists maximum performance, measured by mean average precision, for each language and for words and n-grams. The improvement gains from using n-grams can be

Table 3. Comparison of raw words and n-grams as indexing terms.

	Words		n-grams					
	MAP	α	n	MAP	α	% change	Mean word length	Wilcoxon p-value
Dutch	0.3942	0.5	5	0.4323	0.6	10	5.170	0.0062
English	0.4749	0.6	4	0.4917	0.9	4	4.695	0.2340
Finnish	0.2398	0.9	5	0.3968	0.3	66	7.228	0.0001
French	0.3596	0.6	4	0.4097	0.7	14	4.755	0.0685
German	0.3161	0.9	4	0.4055	0.9	28	5.939	0.0002
Italian	0.3629	0.5	4	0.4051	0.7	12	5.058	0.0253
Spanish	0.4220	0.7	4	0.4678	0.7	11	4.901	0.0367
Swedish	0.3226	0.3	4	0.4312	0.8	34	5.264	0.0001

The table shows values for n and smoothing parameter α that maximize accuracy. Percentage change is relative.

substantial; more than a 25% improvement is seen in German and Swedish, and over 65% with Finnish.

Generally, the improvement when n-grams are used was significant at the 0.05 confidence level, but this was not true for English and French.

4.3. *N-grams and mean word length*

The degree to which n-gram accuracy exceeds raw word accuracy in Table 3 appears to track mean word length. The Pearson correlation coefficient between the improvement in mean average precision when n-grams are used in place of words and the mean word length of a language is 0.935. Our interpretation of this correlation is that (at least in European languages) mean word length is an indicator of morphological complexity, and that raw words work relatively worse with morphologically complex languages.

Mean word length of a language has also been suggested as an indicator of the best n-gram length for that language (e.g., in Lee and Ahn 1996). In fact, bigrams are known to work well for Chinese (Chen et al. 1997) and 6-grams have been shown to perform well in several European languages (Mayfield et al. 2000). Except for the two languages with the longest words, Finnish and German, the mean length is close to 5 characters. Since the CLEF collection comprises news sources, and news articles typically contain many proper names, word length could differ in other genres. These mean word length values do not appear to correlate with optimal n-gram lengths shown in Table 2. We suspect that mean word length is just one of many factors, including morphology and language parsimony, that affect the accuracy of different length of n-gram. We did find that use of either $n = 4$ or $n = 5$ with an α value of 0.5 achieves performance within 5% (relative) of the optimal value achieved in each language by choosing the best combination of length and α value.

Table 4. Best results using 4-grams, 5-grams, or a combination of both 4- and 5-grams.

	4-grams		5-grams		4+5-grams		
	MAP	α	MAP	α	MAP	α	%-change
Dutch	0.4284	0.7	0.4323	0.6	0.4347	0.7	+0.555
Italian	0.4051	0.7	0.4041	0.6	0.4150	0.6	+2.443
Swedish	0.4312	0.8	0.4116	0.6	0.4237	0.6	−1.739

4.4. Use of multiple N-gram lengths

Combination of disparate retrieval techniques has been shown effective in a variety of settings (McNamee et al. 2001b, Savoy 2002). Since 4-grams and 5-grams appear to be the most effective choices for European languages, we examined whether indexing documents using both representations in a combined term space would be beneficial; specifically, we examined Dutch, Italian, and Swedish. Our initial results are not promising; very little difference was evident between the individual and combined term spaces (see Table 4). Furthermore, the technique exacts a performance penalty, doubling the number of postings and increasing both dictionary size and effective query length. It is still possible that a different combination of *n*-gram lengths would improve accuracy (especially for Asian languages).

5. *N*-grams for translingual retrieval

The caliber of available translation resources is a key factor in the accuracy of a system for cross-language information retrieval (McNamee and Mayfield 2002a). Either queries can be translated into the language of documents or documents can be translated into the language of queries (McCarley 1999). The former has received the most attention in the literature and only a few have examined translation of document collections for CLIR (Braschler and Schauble 2000, McCarley 1999, McNamee and Mayfield 2002c, Oard and Hackett 1997). Regardless of which approach is taken, several types of translation resources may be used: machine translation software; bilingual wordlists; and translation equivalents mined from aligned parallel texts. We focus on query-translation in this paper.

5.1. Machine translation

Machine translation (MT) is the simplest translation resource for CLIR systems to use, because it generates (in principle) a well-formed natural language query in the target language. Thus, a monolingual target language retrieval capability requires no changes to accommodate MT-based CLIR.

When *n*-grams are used for indexing, no complications arise from the use of machine translation software. A query is presented to an MT system and the translated query may be mapped into a set of *n*-grams, just as a query that was expressed in the target language would be. A system might choose to remove *n*-grams introduced at the boundaries of words that are passed through untranslated by the MT software, but this technique is unproven.

When other types of translation resources are involved, the process is more involved. We first discuss how parallel collections can be exploited for translingual retrieval. Next, we present a difficulty that is peculiar to *n*-gram retrieval when dictionaries are used. Finally, we demonstrate an ability to achieve serviceable retrieval accuracy relying only on partial cognate matches across related language pairs.

5.2. Parallel collections

Latent Semantic Indexing (LSI) has been applied to the problem of identifying translation candidates for words (Landauer and Littman 1990); they used the Canadian Hansard Corpus, which consists of transcripts of Canadian parliamentary proceedings in both English and French. Subsequently, there has been great interest in exploiting parallel texts for linguistic applications in general and for translation in particular. Several CLEF participants have used parallel corpora in their work (Braschler and Schäuble 2000, Kraaij 2001, McNamee and Mayfield 2002b, Nie et al. 2000). In addition to the Hansard corpus that is available from the Linguistic Data Consortium (LDC), the University of Montreal has mined parallel texts from the Web in French, Italian, and Dutch (Nie et al. 2000). Such texts must first be identified, then aligned at the sentence, or paragraph level (Melamed 2001), and then indexed. Various statistical methods can be used to identify possible translations such as mutual information or a chi-squared test (Och and Ney 2000).

Generally it is difficult to locate parallel texts on the Web (see Resnik 1999). However, it is relatively easy to target sites known to produce a large volume of parallel documents. Taking this approach, JHU/APL has mined the Official Journal of the European Commission over the past two years. The Journal is available in the official EU languages, and is published electronically in Adobe PDF format. We downloaded content nightly, converted the PDF to plain text using a publicly available tool (which does a good job on all languages except Greek), and aligned the documents using the char_align software developed by Church (1993). In this way we obtained about 200 MB of text in each language, which can be aligned with any of the other 10 languages.

Once the aligned collection has been indexed, a statistical translation lexicon can be extracted, mapping words in one language to a set of alternative translations, possibly with translation probabilities for each. Alternatively, the most probable match, or best *k* matches can be extracted (Braschler and Schäuble 2000). Although mappings between words are typically extracted, there is nothing to preclude derivation of other types of mappings when alternative indexing methods are used. For example, if *n*-grams are used, statistical relationships between *n*-grams in one language and those of a different language can be identified. The lengths of *n* may even differ. Or, if words are used in one language and *n*-grams in

Table 5. (a) Highest ranking Spanish 5-grams corresponding to a variety of English 6-grams. (b) Sample English words with the highest ranking Spanish word and 5-gram translation for each.

(a)		(b)		
English 6-gram	Spanish 5-gram	English word	Spanish translation	Spanish 5-gram
in-the	en-la	Any	calquier	lquie
-in-th	en-la	Insurance	seguro	eguro
new-op	-nuev	Years	años	-años
-the-p	l-pro	Mail	correo	rreo-
-witho	-sin-	New	nuevo	-nuev
servic	ervic	Area	zona	zona-
n-which	n-que	Neighbouring	vecino	-veci
n-area	-zona	Outgoing	saliente	-sali

another, 'n-gram translations' of each word can be created. We used this approach to map English words to Chinese bigrams for English to Chinese CLIR (McNamee et al. 2001b).

Our method for producing translations of a term t starts with finding the set S of all source language documents that contain term t. We then select the set S' of target language documents that are translations of documents in S. We count the frequency of each word that occurs at least once in S'. We then identify terms that occur much more frequently in S' than they do in the target language collection as a whole; the best candidate translation for t, t', is the term exhibiting the largest such difference. The process took over a CPU-day for word-to-word mappings and several days when n-grams were used. Table 5 contains sample mappings between English 6-grams or raw words, and Spanish 5-grams.

To investigate bilingual retrieval using this method, we created several lexicons that could be used for English-to-Spanish retrieval and evaluated the accuracy of each using the CLEF 2002 English topics to search the Spanish sub-collection. No pre-translation query expansion or post-translation relevance feedback was applied to these runs, though both techniques generally improve cross-language retrieval accuracy (Ballesteros and Croft 1997, 1998, McNamee and Mayfield 2002a). Results using each resource are shown in Table 6. Raw words and 5-grams achieve comparable accuracy for monolingual retrieval; the mean average precision when words were used was 0.4220, which increases slightly to 0.4479 (+6.13%) with 5-grams. However, the difference between the two increased for bilingual retrieval. When English words were translated to Spanish words, a mean average precision of 0.3071 was observed. The corresponding run mapping English 5-grams to Spanish 5-grams has mean average precision of 0.3539, a 15.3% relative improvement.

One of our goals for this experiment is to demonstrate that an end-to-end system for cross-language information retrieval can avoid language-specific processing, even for the translation component. In fact, in this example, the best bilingual accuracy was obtained in exactly this fashion.

Table 6. Various methods for bilingual retrieval using parallel corpora. In the bilingual runs, only a single translation was selected for each source language term.

	Source terms	Target terms	Target α	MAP	Prec. @ 10 docs
Monolingual	EN words	EN words	0.6	0.4749	0.4024
Monolingual	EN 5-grams	EN 5-grams	0.8	0.4718	0.4333
Monolingual	ES words	ES words	0.7	0.4220	0.5340
Monolingual	ES 5-grams	ES 5-grams	0.5	0.4479	0.5580
Bilingual	EN words	ES words	0.7	0.3071	0.3900
Bilingual	EN words	ES 5-grams	0.5	0.2471	0.3320
Bilingual	EN 5-grams	ES words	0.7	0.1415	0.2360
Bilingual	EN 5-grams	ES 5-grams	0.5	0.3539	0.4520

5.3. *Bilingual dictionaries*

Machine translation and parallel collection-based translation work smoothly with n-gram representations of the target language. Their output naturally contains word-spanning n-grams because it is drawn from full target language sentences. In contrast translation dictionaries support only word-internal n-grams (unless the dictionary contains target language phrases). In Section 4 we showed that word-spanning n-grams may approximate multiword phrases such as “white house”. Use of a bilingual dictionary eliminates the availability of n-grams to capture this phrasal information, and this might degrade retrieval accuracy, if phrasal information contributes to that accuracy (see Pirkola et al. 2001).

To quantify the effect of losing word-spanning n-grams, it is insufficient to compare the accuracy of a dictionary-based method against that of another method; the results would conflate the effects of a lack of word-spanning n-grams with the effects of differences in lexical coverage between the two translation resources. Instead, we compare monolingual retrieval accuracy for word-spanning n-grams and word-internal n-grams for the same test collection. Although this is a monolingual evaluation, its primary significance is for dictionary-based cross-language retrieval. (Word-internal n-grams are readily available in monolingual retrieval, so restricting indexing terms to word-internal n-grams would only be useful as a term reduction technique.) We examined Dutch, English, and Finnish in this fashion using the CLEF 2002 test suite and using 5-grams in each language. Table 7 compares retrieval accuracy for spanning and non-spanning n-grams for these three languages. Our experiments show that whether the elimination of word-spanning n-grams degrades accuracy varies with language. Accuracy suffered slightly in Dutch and Finnish; the decrease was significant in Dutch ($p = 0.034$, Wilcoxon test). We had anticipated that ignoring word-spanning n-grams would decrease accuracy by eliminating surrogate phrasal information. Since the loss appears to be minor, we conclude that n-grams are not contraindicated for dictionary-based cross-language retrieval.

Table 7. Comparing retrieval using spanning 5-grams and word-internal 5-grams. Only the Dutch result is statistically significant.

	5-grams		Word-internal 5-grams		
	MAP	α	MAP	α	%-change
Dutch	0.4323	0.6	0.4141	0.7	−4.210
English	0.4718	0.8	0.4746	0.8	+0.593
Finnish	0.3968	0.3	0.3888	0.2	−2.016

This is consistent with studies of word-based phrasal indexing (Voorhees and Harman 1999).

5.4. *No translation*

Translation resource availability varies widely across language pairs. For some language pairs, resources may be entirely unavailable. An alternative to query translation in such cases is to leave the query untranslated. This approach seems peculiar at the outset, but surprisingly good results can be obtained without translation when European languages are used. The reason is that cognates abound across the CLEF languages.

Plausible scenarios exist in which the no-translation approach might be useful. For example, suppose we want to search a collection of Galician documents using a query expressed in English. We might be able to obtain reasonable accuracy by translating the English query into Portuguese (for which good resources are available), and relying only on cognate matches to move from Portuguese to Galician.

Buckley et al. (1998) examined retrieval of French from English queries by treating English as misspelled French. Natural English/French cognates were augmented by rules to 'spell-correct' English words to French words of similar spelling. They estimated that 30% of non-stopwords could so be transformed automatically. Unpredictably, this approach was enormously successful relative to other cross-language approaches of the time, achieving the highest scoring automatic cross-language run at the TREC-6 evaluation (mean average precision of 24%, which was about 60% of monolingual). The authors point out that this technique is only useful for related languages.

The Buckley approach demands language-specific tools for spelling correction. We find that similar techniques are effective without language-specific resources. Using cognate matches drawn from raw words is relatively ineffective. N-gram tokenization greatly improves the accuracy of the no-translation approach, because the number of matching n-grams across related languages is significantly greater than the number of words that match exactly. Simply using n-gram matches instead of word cognates can double the efficacy of this approach. The use of pre-translation query expansion may result in even better performance but we have not explored this. Figure 2 shows the accuracy of the no-translation approach for a variety of language pairs.

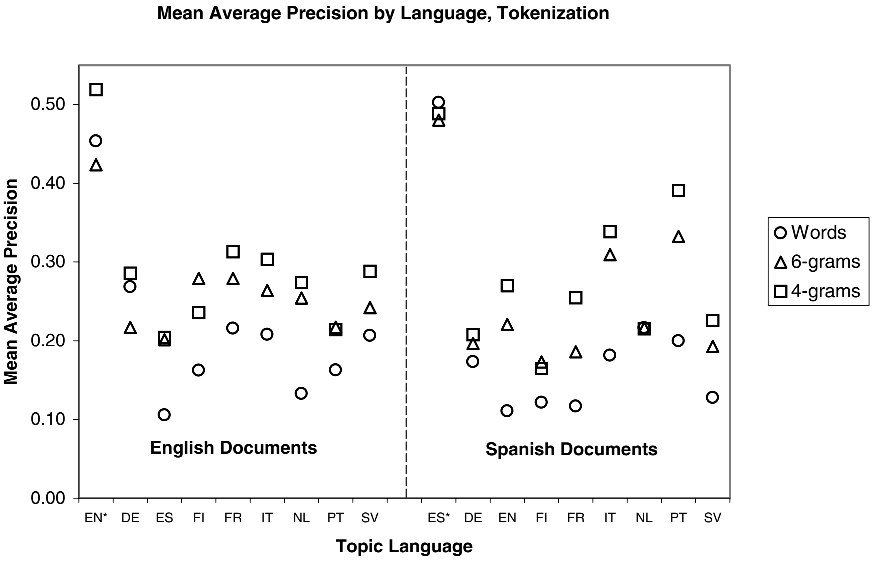


Figure 2. Using partial cognate matches for CLIR.

Since the no-translation approach works only when the source and target languages are closely related, this leaves open the question of how similar two languages must be for this technique to be effective. We adapted a method by Benedetto et al. (2002) to calculate the relatedness of two languages, with the goal of determining whether the accuracy of no-translation retrieval can be predicted given a language pair.

Benedetto and colleagues describe a method for computing the similarity of arbitrary text sequences, and illustrate the technique for both author identification and language similarity estimation. To estimate the similarity between two languages the authors made use of a widely used compression algorithm, *gzip*, and a document that is available in many languages. The technique divides each version of the document into two pieces, a larger training piece denoted by uppercase, and a smaller residual piece denoted by lowercase. It computes how well a second piece can be encoded using a compression technique based on statistics learned from a first portion. For a well-trained model, the second portion should be compactly encoded if it is much like the first part. Let ΔAb indicate the difference between the length of the compressed versions of the conjoined content Ab and A alone. For example, $\Delta ENfr$ would indicate the difference between the size of the compressed file created by appending the residual French text to the larger English training portion and the size of the compressed English training portion alone. Using such differences, an estimate of the distance between two languages can be calculated as

$$d = (\Delta Ab - \Delta Bb)/\Delta Bb + (\Delta Ba - \Delta Aa)/\Delta Aa$$

This unitless value can be computed for each language pair. Using the above method applied to the Universal Declaration of Human Rights (United Nations, 1948), we obtain the following symmetric matrix:

	DE	EN	ES	FI	FR	IT	NL	PT	SV
DE	0.000	0.548	0.548	0.570	0.547	0.574	0.486	0.555	0.494
EN	0.548	0.000	0.450	0.586	0.457	0.486	0.524	0.461	0.513
ES	0.548	0.450	0.000	0.549	0.425	0.429	0.529	0.315	0.515
FI	0.570	0.586	0.550	0.000	0.546	0.613	0.557	0.572	0.529
FR	0.547	0.457	0.425	0.546	0.000	0.453	0.542	0.456	0.520
IT	0.574	0.486	0.429	0.613	0.453	0.000	0.555	0.444	0.565
NL	0.486	0.524	0.529	0.557	0.542	0.555	0.000	0.546	0.493
PT	0.555	0.461	0.315	0.572	0.456	0.444	0.549	0.000	0.535
SV	0.494	0.513	0.515	0.529	0.520	0.565	0.493	0.535	0.000

Using Spanish as an example, figure 3 shows that the Benedetto language pair distance between Spanish and each of the other languages is predictive of mean average precision for 6-gram no-translation retrieval with a Pearson correlation coefficient of 0.94.

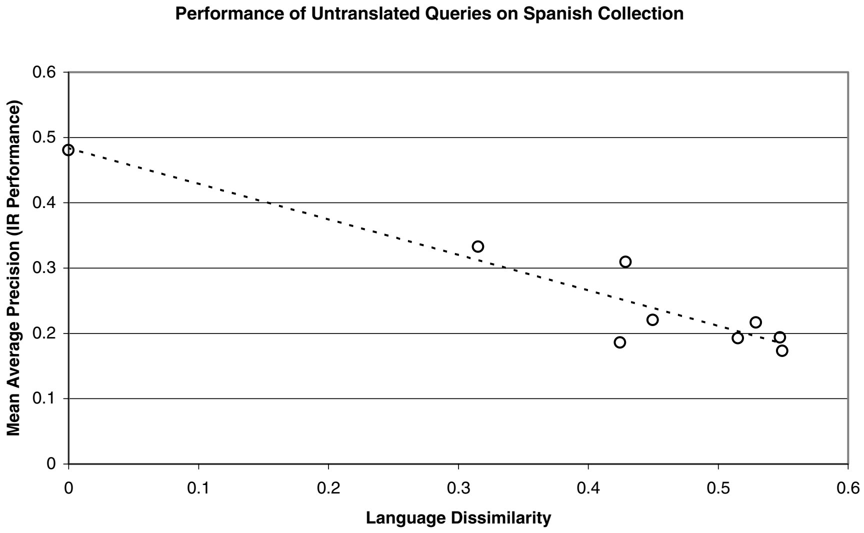


Figure 3. Language similarity and application of partial cognate matches for CLIR. There is a correlation between the closeness of Spanish to other languages and the accuracy of untranslated queries in cross-language text retrieval. Points represent retrieval accuracy over Spanish documents from untranslated queries in nine European languages.

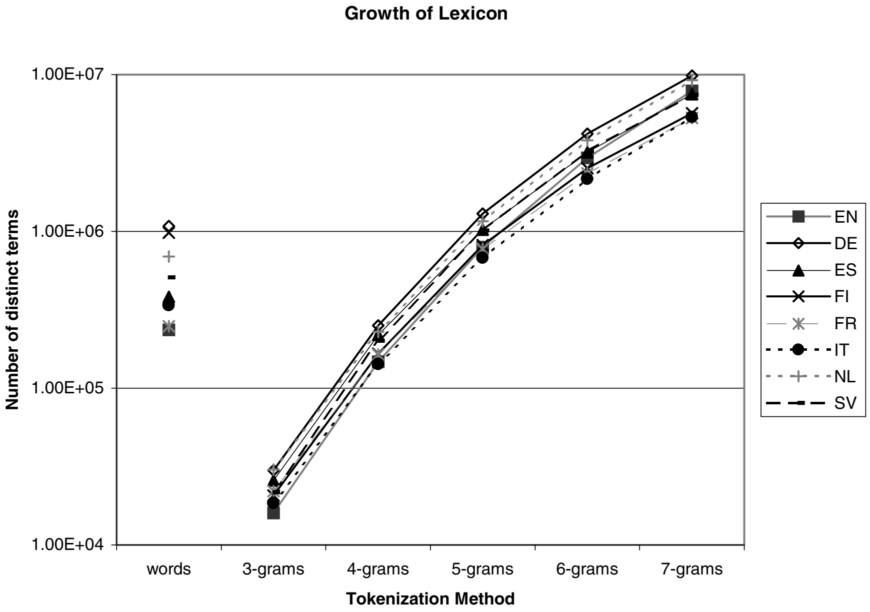


Figure 4. Lexicon growth for words and character n-grams.

6. Performance

While we have demonstrated that the use of n-grams can improve retrieval accuracy, we have not discussed the performance ramifications of n-gram indexing. While early work looked favorably on short n-grams, which reduce the dictionary size, we find that longer-length n-grams provide higher accuracy. Dictionary size grows rapidly with n, but the worst-case exponential behavior $|alphabet|^n$ is not observed. We show the number of distinct n-grams as a function of n-gram length in figure 4.

Although the size of the lexicon remains an important consideration (particularly for long n-grams), a more significant consequence of n-gram indexing is an increase in index creation and query processing times. Query execution times are often dominated by disk accesses, so it is important to know how the size of the inverted index grows with alternative tokenization schemes; this is also important for assessing the increased storage requirements of n-grams. In figure 5 the number of entries in the inverted file for the Spanish CLEF 2002 collection is plotted for different tokenization methods (assuming that the inverted file includes only document IDs and not term positions within the document). Even for 3-grams, the total number of entries is much greater than for words, because many 3-grams are present in a large proportion of documents. In Table 8 we report the mean length of a postings list in each language (note though that the distributions are skewed by stopwords and 'stop' n-grams).

To gain a better understanding of the distribution of terms using the different tokenization methods, we can plot the number of terms that occur in a given range of document counts.

Table 8. Average length of a postings list in the inverted file.

	Words	3-grams	4-grams	5-grams	6-grams	7-grams
Dutch	53.04	4678.03	1002.54	250.29	86.27	38.31
English	127.87	6802.70	1250.62	292.74	89.45	36.02
Finnish	10.41	2041.33	396.49	96.80	34.13	15.86
French	69.48	3038.48	641.53	170.05	62.86	29.64
German	39.69	5979.36	1089.24	259.00	89.04	40.26
Italian	64.59	4189.12	912.94	239.82	83.59	35.50
Spanish	98.41	6198.36	1111.75	286.42	101.99	45.99
Swedish	34.82	3762.58	572.14	130.99	44.15	20.11

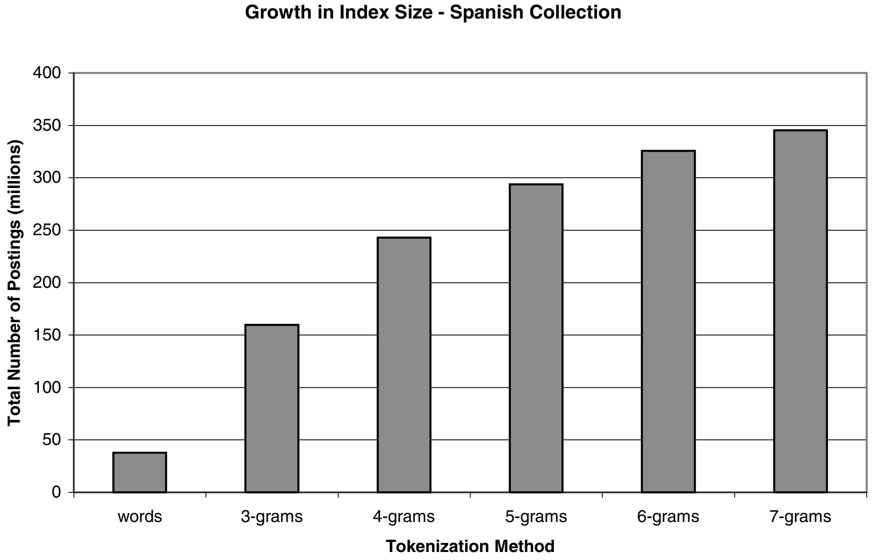


Figure 5. Size of inverted file for various tokenization methods.

Figure 6 shows the distribution of terms that occur in a specified number of documents for the CLEF 2002 Spanish collection, by tokenization scheme. The data points at the far left of the plot are those that occur in only a single document; the rightmost points indicate the number of terms that occur in 50% or more of the documents.

Finally, in Table 9 we report on query lengths (in terms) and HAIRCUT query execution times (wall time in seconds) for the Spanish topics. There is a hefty increase in query processing times when n-grams are used; retrieval with 4-grams and 5-grams is ten times slower. A portion of this increase in processing time may be an artifact of our current implementation, but the trend is unmistakable. Interestingly the response time decreases with longer n-grams. We interpret this to be due in part to the slightly lower number of

Table 9. Execution times (seconds) and query lengths (in terms) for the CLEF 2002 Spanish topics.

	Total time	Increase vs. words	Mean time	Std. dev.	Mean length	Std. dev.
Words	174.8		3.5	2.4	9.6	3.3
3	724.8	4.15	14.5	6.2	67.1	21.4
4	1860.5	10.64	37.2	14.2	65.3	22.0
5	1851.7	10.59	37.0	16.6	61.4	21.6
6	1531.4	8.76	30.6	16.1	56.9	21.0
7	1124.1	6.43	22.5	13.4	52.0	20.1

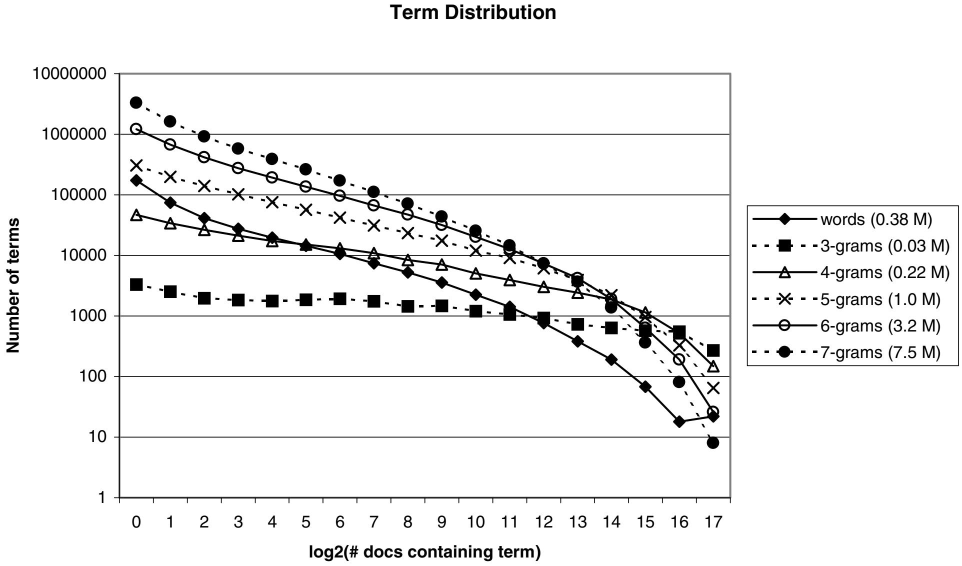


Figure 6. Term distributions in Spanish for multiple indexing units.

query terms, and in part to the smaller average postings list length found with larger *n* (see Table 8).

The increase in response time when *n*-grams are used suggests that *n*-grams should be used cautiously when a rapid response is required. However, we have paid scant attention to these sorts of performance issues, and our current implementation may well be biased towards words. Furthermore, aggressive pruning techniques can reduce the size of the index without harming accuracy (Pearce and Rye 1998, Carmel et al. 2001).

7. Conclusions

N-grams are a viable alternative to words as indexing terms in information retrieval. In our tests, *N*-grams provide higher accuracy than a strawman system using raw words as indexing terms in each of the eight languages of the CLEF collection; languages with greater mean word length fared relatively better with n-grams than with words. An improvement of more than 25% in mean average precision was observed in the Finnish, German, and Swedish languages. N-gram lengths of 4 and 5 worked well for all languages and selection of either length with an α value of 0.5 always yielded performance within 5% of the optimal setting.

N-grams also work well for cross-language retrieval, under all four types of cross-language mapping: machine translation, parallel corpora, bilingual dictionaries, and no translation. Parallel corpora allow the source language to be tokenized differently from the target language, supporting a mixed mode in which words are used only for languages with readily available lexical resources. Bilingual dictionaries will not support word-spanning n-grams, but our tests indicate that this should not be particularly detrimental to retrieval accuracy. When no translation resources are available, n-gram tokenization can still lead to astonishingly good cross-language retrieval performance for related languages.

Performance is an issue when indexing using n-grams; a significant increase is seen in the size of both the dictionary and the postings list. However, the penalty paid is heavily dependent on implementation; we have not explored these engineering issues in depth.

Given these results, a researcher interested in developing a monolingual or translingual retrieval capability for a language that lacks significant language-specific resources should strongly consider using n-gram tokenization.

References

Ballasteros L and Croft WB (1997) Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, pp. 84–91.

Ballasteros L and Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 64–71.

Benedetto D, Caglioti E and Loreto V (2002) Language Trees and Zipping. Physical Review Letters, 88.

Braschler M and Schäuble P (2000) Experiments with the eurospider retrieval system for CLEF 2000. In: Peters C, Ed. Proceedings of the First Cross-Language Evaluation Forum, pp. 140–149.

Buckley C, Mitra M, Walz J and Cardie C (1998), Using clustering and super concepts within SMART: TREC-6. In: Voorhees EM and Harman DK, Eds. Proceedings of the Sixth Text REtrieval Conference (TREC-6), NIST Special Publication 500-240, pp. 107–124.

Carmel D, Cohen D, Fagin R, Farchi E, Herscovici M, Maarek Y and Soffer A (2001) Static index pruning for information retrieval systems. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–50.

Cavnar WB (1994) Using an N-gram-based document representation with a vector processing retrieval model. In: Harman DK, Ed. Proceedings of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226, pp. 269–278.

Cavnar WB and Trenkle JM (1994) N-Gram based text categorization. In: Proceedings of the Third Symposium on Document Analysis and Information Retrieval, pp. 161–169.

Chen A, He J, Xu L, Gey F and Meggs J (1997) Chinese text retrieval without using a dictionary. SIGIR, 42–49.

Church KW (1993) Char_align: A program for aligning parallel texts at the character level. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, pp. 1–8.

Cohen JD (1995) Highlights: Language- and domain-independent automatic indexing terms for abstracting. Journal of the American Society for Information Science, 46:162–174.

Comlekoglu FM (1990) Optimizing a text retrieval system utilizing N-gram indexing. Ph.D Thesis, George Washington University.

Damashek M (1995) Gauging similarity with n-grams: Language-independent categorization of text. Science, 267:843–848.

D'Amore RJ and Mah CP (1985) One-time complete indexing of text: Theory and practice. In: Proceedings of the 8th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-85), pp. 155–164.

De Heer T (1974) Experiments with syntactic traces in information retrieval. Information Storage and Retrieval, 10:133–144.

De Heer T (1982) The application of the concept of homeosemy to natural language information retrieval. Information Processing & Management, 18:229–236.

Harman D (1992) Relevance feedback revisited. In: Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-92), pp. 1–10.

Harman D et al. (1995) Performance of text retrieval systems. Science, 268:1417–1418.

Hiemstra D (2000) Using language models for information retrieval. Ph.D. Thesis, Center for Telematics and Information Technology, The Netherlands.

Jelinek F and Mercer R (1980) Interpolated estimation of Markov source parameters from sparse data. In: Gelsema ES and Kanal LN, Eds. Pattern Recognition in Practice, North Holland, pp. 381–402.

Kraaij W (2001) TNO at CLEF-2001: Comparing translation resources. In: Peters C et al., Eds. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF-2001).

Landauer TK and Littman ML (1990) Fully automated cross-language document retrieval using latent semantic indexing. In: Proceedings of the 6th Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp. 31–38.

Lee JH and Ahn JS (1996) Using N-grams for Korean text retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 216–224.

Mah CP and D'Amore RJ (1983) Complete statistical indexing of text by overlapping word fragments. ACM SIGIR Forum, 17(3):6–16.

Mayfield J, McNamee P and Piatko C (2000) The JHU/APL HAIRCUT system at TREC-8. In: Voorhees EM and Harman DK, Eds. Proceedings of the Eighth Text REtrieval Conference (TREC-8), NIST Special Publication 500-246, Gaithersburg, Maryland, pp. 445–452.

McCarley S. (1999) Should we translate the documents or the queries in cross-language information retrieval. In: Proceedings of ACL.

McNamee P, Mayfield J and Piatko C (2001a) A language-independent approach to European text retrieval. In: Peters C Ed. Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF-2000 Workshop, Lecture Notes in Computer Science 2069, Springer, Lisbon, Portugal, pp. 129–139.

McNamee P, Mayfield J and Piatko C (2001b) The HAIRCUT system at TREC-9. In: Voorhees EM and Harman DK, Eds. Proceedings of the Ninth Text REtrieval Conference (TREC-9), NIST Special Publication 500-249, Gaithersburg, Maryland, pp. 273–279.

McNamee P and Mayfield J (2002a) Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 159–166.

McNamee P and Mayfield J (2002b) JHU/APL experiments at CLEF-2001: Translation resoruces and score normalization. In: Peters C et al. Eds. Evaluation of Cross-Language Information Retrieval Systems: Second Workshop of the Cross-Language Evaluation Forum (CLEF-2001), Darmstadt, Germany, pp. 193–208.

McNamee P and Mayfield J (2002c) Scalable multilingual information access. In: Working Notes of the CLEF 2002 Workshop, Rome, Italy, pp. 133–140.

Melamed ID (2001) Empirical Methods for Exploiting Parallel Texts. MIT Press, Cambridge, MA.

Mihalcea R and Nastase V (2002) Letter level learning for language independent diacritics restoration. In: Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002), pp. 105–111.

Miller D, Leek T and Schwartz R (1999) A hidden Markov model information retrieval system. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, California, pp. 214–221.

Nie J-Y, Simard M and Foster G (2000) Multilingual information retrieval based on parallel texts from the web. In: Proceedings of the CLEF-2000 Workshop, Lecture Notes in Computer Science 2069, Springer, Lisbon, Portugal, pp. 188–201.

Oard DW and Hackett P (1997) Document translation for cross-language text retrieval at the University of Maryland. In: Proceedings of the Sixth Text REtrieval Conference (TREC-6), pp. 687–696.

Oard DW, Levow G and Cabezas CI (2001) CLEF experiments at Maryland: Statistical stemming and backoff translation. In: Peters C, Ed. Proceedings of the First Cross-Language Evaluation Forum, pp. 176–187.

Och FJ and Ney H (2000) Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440–447.

Ogawa Y and Matsuda T (1997) Overlapping statistical word indexing: A new indexing method for Japanese text. In: Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR-97), pp. 226–234.

Pearce C and Nicholas C (1996) TELLTALE: Experiments in a dynamic hypertext environment for degraded and multilingual data. Journal for the American Society for Information Science, 47:236–275.

Pearce C and Rye W (1998) N-gram term weighting: A comparative analysis. National Security Agency Technical Report, TR-R52-001-98.

Peters C and Braschler M (this volume), Manuscript in preparation.

Pirkola A, Hedlund T, Keskusalo H and Järvelin K (2001) Dictionary-based cross-language information retrieval: Problems, methods, and research findings. Information Retrieval, 4:209–230.

Ponte JM and Croft WB (1998) A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 275–281.

Porter MF (1980) An algorithm for suffix stripping. Program, 14:130–137.

Porter MF (2001) Snowball: A Language for Stemming Algorithms. http://snowball.tartarus.org/texts/introduction.html (visited 13 March 2003).

Robertson SE, Walker S and Beaulieu M (1999) Okapi and TREC-7: Automatic ad hoc, filtering, vlc, and interactive. In: Voorhees EM and Harman DK, Eds. Proceedings of the 7th Text REtrieval Conference (TREC-7), August 1999, NIST Special Publication 500-242, pp. 253–264.

Resnik P (1999) Mining the web for bilingual text. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 527–534.

Salton G and Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5):513–523.

Savoy J (2002) Report on CLEF 2002 experiments: Combining multiple sources of evidence. In: Working Notes for the CLEF 2002 Workshop, pp. 31–46.

Savoy J (2003) Cross-language information retrieval: Experiments based on CLEF 2000 corpora. Information Processing and Management, 39(1):75–115.

Shannon C (1948) A mathematical theory of communication. Bell System Technical Journal, 27:379–423 and 623–656.

Shannon C ($\sim$1980) Scientific aspects of juggling. In: Sloane NJA and Wyner AD, Eds. (1993) Claude Elwood Shannon: Collected Papers, IEEE Press.

Teufel B (1988) Natural language documents: Indexing and retrieval in an information system. In: Proceedings of the 9th International Conference on Information Systems, Minneapolis, Minnesota, pp. 193–201.

United Nations (no date). Universal Declaration of Human Rights, http://www.unhchr.ch/udhr/ (visited October 28th, 2002).

Voorhees EM and Harman DK (1999) Overview of the seventh Text REtrieval Conference (TREC-7). In: Voorhees EM and Harman DK, Eds. The Seventh Text REtrieval Conference (TREC-7). NIST Special Publication 500-242.

Willett P (1979) Document retrieval experiments using indexing vocabularies of varying size. II. Hashing, truncation, digram and trigram encoding of index terms. Journal of Documentation, 35:296–305.

Witten IH, Moffat A and Bell TC (1999) Managing Gigabytes—Compressing and Indexing Documents and Images, 2nd ed., Morgan Kaufmann Publishers.

Zamora EM, Pollock JJ and Zamora A (1981) The use of trigram analysis for spelling error detection. Information Processing and Management, 17:305–316.

Zhai C and Lafferty J (2001) A study of smoothing methods for language models applied to ad hoc information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 334–342.